Dataverse, an open and collaborative source research data repository software

Baptiste Rouxel, Alina Danciu, Jim Meyers, Geneviève Michaud, Tom Villette

SciencesPo
CENTRE DE DONNÉES SOCIO-POLITIQUES

DDI

The Dataverse Project

cnrs

Dataverse, an open and collaborative source research data repository software

Alina Danciu[1], Geneviève Michaud[1], Jim Myers[2], Baptiste Rouxel[1], Tom Villette[1]

[1]*Center for Socio-Political Data (CDSP)*

[2] *Global Dataverse Community Consortium (GDCC)*

-

“Dataverse is an open source software platform for sharing, finding, citing, and preserving research data (developed by the Data Science and Products team at the Institute for Quantitative Social Science and the Dataverse community).”

Source : https://github.com/IQSS/dataverse

What?

75 installations worldwide

Software installation, which then hosts multiple virtual archives called Dataverse collections. Each Dataverse collection contains datasets, and each dataset contains descriptive metadata and data files (including documentation and code that accompany the data). As an organizing method, Dataverse collections may also contain other Dataverse collections.

Why?

Dataverse is a good solution for networked institutions

Single Sign On using institutional credentials
Can harvest and be harvested
General and Dataverse-level branding capabilities

Dataverse supports discoverability and persistence

Powerful search engine and filters
Good Google referencing
Persistent identifiers

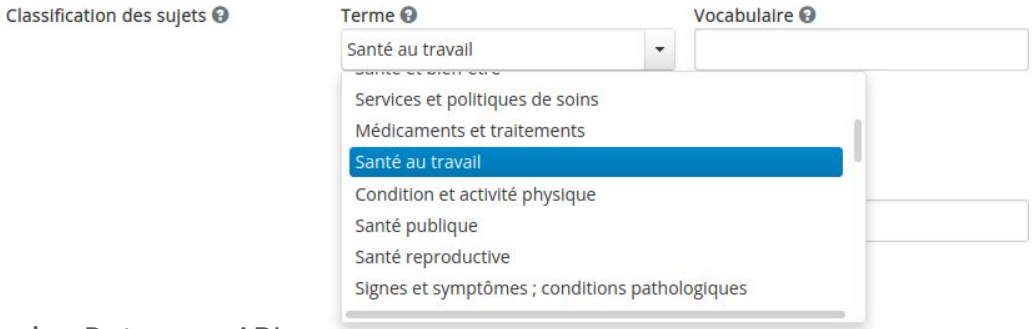
Danciu, Alina, & Michaud, Geneviève. (2020, December 1). DDI, Dataverse and Colectica: our data management combo. 12th Annual European DDI User Conference (EDDI20). Zenodo. https://doi.org/10.5281/zenodo.4299083

What for?

Nesstar => Dataverse

Harmonising metadata (institution names - in-house + ROR and ISNI, authors, keywords -ELSST...)

Implementing DDI controlled vocabularies



How?

In-house development of several scripts, some using Dataverse APIs:

- metadata harmonisation & conformity to CESSDA Metadata Model (CMM)
- migrating metadata from Nesstar to Dataverse

Installation and configuration of a Dataverse instance enabling:

- controlled vocabularies (according to the CMM, DDI CVs)
- multilingual UI (including CVs translation)
- OAI-PMH harvesting

data.sciencespo repository

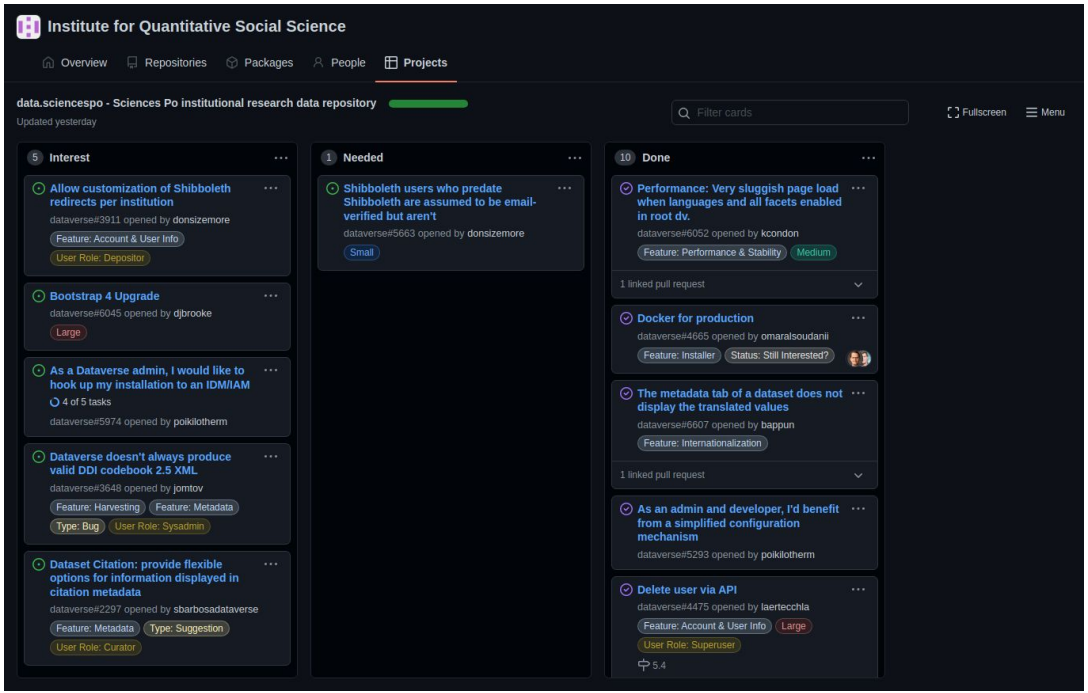
CDSP data bank & institutional self deposit repository released January 2020:

- 7 custom CVs
- 300+ metadata items harmonised & migrated
- EN | FR user interface
  - including CVs translations
- custom homepage and style
- harvested by:
  - CESSDA DATA Catalogue (select French UI)
  - french SSH portal Isidore
  - European OpenAIRE portal

Dataverse communities

- Other groups working on similar issues (PyDataverse, SuperDADA etc)
- CDSP's contributions to IQSS DV codebase (pull requests)
- Opening of several issues aimed at OAI-PMH endpoints
    - harvesting by the CESSDA Data Catalogue
- Decision to sponsor a list of key issues (GDCC contract and collaboration)
    - tackling **concerns shared among users in the DDI community**

https://github.com/orgs/IQSS/projects/12



Internationalization in Dataverse

Viewing in multiple languages

- User Interface translations
- Translations of metadata fields with controlled vocabularies

Issues raised by Sciences Po re: Content Creation/Machine Harvesting:

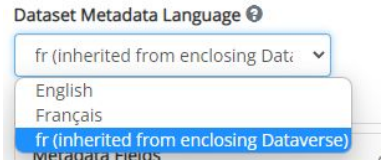
- Metadata field to specify the language(s) used in data files, but
- No way to indicate the language used to enter metadata for a dataset
- Metadata exports (including the DDI export used for harvesting by CESSDA) do not indicate the language used

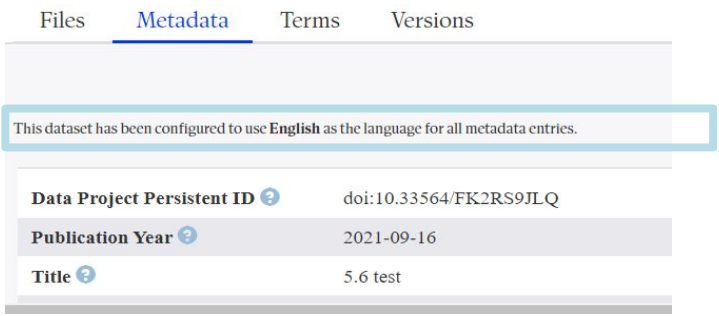
Specifying the Metadata Language

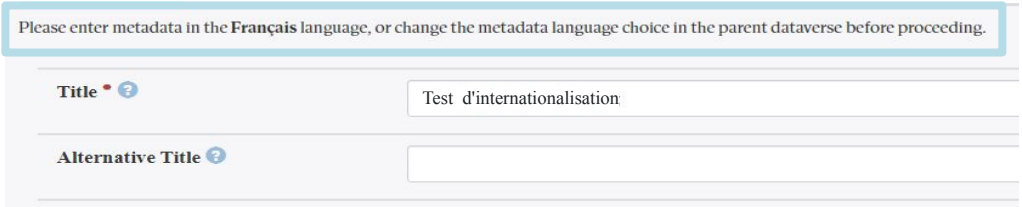
Admins enable/specify the allowed languages

Sub-collections are configured to indicate the metadata language allowed for new datasets

Dataset creators/editors are prompted to use the configured language







Exporting Language Information

Metadata exports include metadata language, i.e. DDI:

Overall xml:lang attribute

```
<codeBook xmlns="ddi:codebook:2_5"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="ddi:codebook:2_5
https://ddialliance.org/Specification/DDI-Codebook/2.5/XMLSchema/codebook.xsd"
version="2.5" xml:lang="fr">
```

Individual xml:lang attributes on specific fields

```
<abstract xml:lang="fr">TEST</abstract>
```

Controlled values exported in site default language and dataset's metadata language

```
<sumDscr>
  <anlyUnit xml:lang="en">Family: Household family</anlyUnit>
  <anlyUnit xml:lang="en">Household</anlyUnit>
  <anlyUnit xml:lang="fr">Famille : ménage</anlyUnit>
  <anlyUnit xml:lang="fr">Ménage</anlyUnit>
</sumDscr>
```

* Thanks for the online CESSDA validator! - used to identify where language information was needed/recommended

Community Processes

The Dataverse® Project

Open Source on GitHub

Community-contributed issues and code contributions

Community contributions to language translations

GDCC - Global Dataverse Community Consortium

Prioritizing work of interest across the community

Contracting to enable members to engage experienced developers/admins on their priorities

Dataverse 5.7+

New functionality for describing datasets/supporting harvesting by CESSDA in multiple languages thanks to Sciences Po!